Early Chronic Illness Detection

Vickey Kumar (2021299), Suyash Kumar (2021293), Aditya Jain (2021305), Satyam (2021285), Satyam Pandey (2022463)

October 23, 2024

1 Abstract

Globally, an estimated 3.8% of the population experiences depression, including 5% of adults and 5.7% of adults older than 60 years. India's suicide rate is among the highest globally, with an estimated 21.1 suicides per 100,000 people, highlighting the severe impact of untreated mental health issues. Depression is a leading cause of disability worldwide, contributing significantly to the global burden of disease, and in India, it is pushing around 20% of households into poverty due to high healthcare costs. Timely detection of depression is crucial for preventing severe mental health conditions and enabling more cost-effective treatments.

2 Introduction

This project aims to develop a cost-effective and holistic solution for depression detection by incorporating socio-economic, health, and lifestyle data. By analyzing this comprehensive dataset, we seek to improve prediction accuracy and provide deeper insights into the relationship between these factors and chronic illness, enabling more accurate early detection and intervention strategies.

3 Literature Review

3.1 Depression Detection Using Machine Learning (Chauhan et al. 2023)

This study establishes a relationship between stress-related factors and depression detection using machine learning, focusing on identifying individuals with high stress levels who may not show visible symptoms of depression [2]

3.2 Neuroimaging-Based Detection (Fu et al. 2018)

Fu et al. utilized machine learning on neuroimaging data to detect depression with 86% accuracy by identifying brain structural changes, using cross-validation for robust results [3].

3.3 A comparative study of different classifiers for detecting depression from spontaneous speech

This research employed Support Vector Machines (SVM) and Natural Language Processing (NLP) to analyze speech features, such as slower rates and longer pauses, achieving 80% accuracy in predicting depression [1]

4 Dataset Description

4.1 Data Collection

The primary source of our dataset is a Kaggle Dataset authored by Anthony Therrien, designed to facilitate the analysis of individuals' health, lifestyle, and socio-economic factors in relation to depression detection.

4.2 Data Features

Important Demographic Information and Personal Attributes:

- **Education Level:** Associate Degree (19.33%), Bachelor's Degree (32.71%), High School (22.81%), Master's Degree (21.15%).
- **Marital Status:** Divorced (7.91%), Married (58.11%), Single (17.43%), Widowed (16.55%).
- **Smoking Status:** Former Smoker (26.67%), Non-smoker (60.72%), Smoker (12.61%).
- **Physical Activity Level:** Active (26.68%), Moderate (31.64%), Sedentary (41.68%).
- **Alcohol Consumption:** Low (22.72%), Moderate (46.80%), High (30.48%).
- **Dietary Habits:** Ranges from healthy to unhealthy eating patterns.
- **Sleep Patterns:** Fair (41.73%), Good (39.48%), Poor (18.79%).
- **Employment Status:** Describes whether the individual is employed or unemployed.
- **Income:** Income ranges from $0.41 to $209,995.22, with an average income of $50,661.71.

- **History of Mental Illness:** Yes (45.59%), No (54.41%).
- **History of Substance Abuse:** Indicates whether the individual has a history of substance abuse.
- **Family History of Depression:** Yes (26.89%), No (73.11%).
- **Chronic Medical Conditions:** Yes (32.92%), No (67.08%).

4.3 Data Distribution

The dataset consists of 413,768 entries and contains 16 columns, including a label for Chronic Disease, which is categorized as 'Yes' or 'No'.

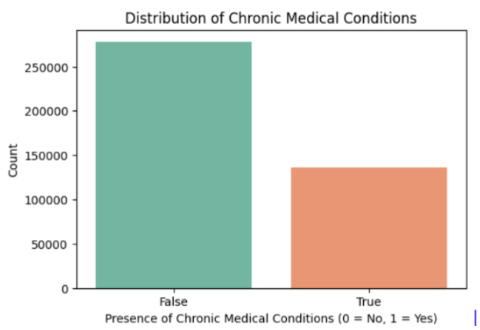


Figure 1: The distribution of chronic medical conditions shows a higher count of individuals labeled 'No'67% compared to those labeled 'Yes'33%

EDA Pattern Detection in Data

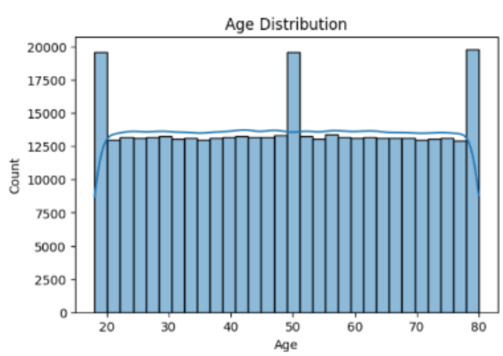


Figure 2: The distribution of chronic medical conditions shows a higher count of individuals labeled 'No'67% compared to those labeled 'Yes'33%

4.4 Data Pre-Processing

- Class Imbalance - occurs as "No" (those without chronic conditions) has significantly more instances than "Yes"(those with chronic conditions) [1]. So to balance the dataset we use SMOTE

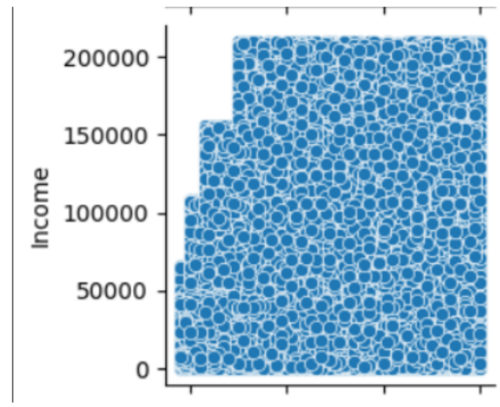


Figure 3: Age distribution shows peaks at 20, 50, and 80, suggesting targeted survey samples.] Relationships between age, income, lifestyle, and chronic conditions are likely non-linear in health data.

- Standardization - of the dataset to improve the model predictions and accuracy later on however minor tweaking is expected to be needed as we move on further into the project.
- Encoding was needed for the varchar values (Categorical Value) ie: Marital Status, Education Level, Smoking Status, Employment Status and Alcohol Consumption etc in the dataset so we have used one-hot encoding for it
- Missing values were printed however the dataset has no NaN or missing values hence not needed to be removed/ replaced with the average.
- Non Linear Relationships : Some feature like age [2] and income shows non linear relationships so Feature engineering can improve the predictive power of the model by helping it better understand the relationships between variables that are not linearly related.

5 Methodology

5.1 Logistic Regression with Batch Gradient Descent (BGD)

- **Parameters:** `solver='lbfgs', max_iter=1000, class_weight='balanced', random_state=42`.
- Applied to handle class imbalance using class weights and trained on scaled data for better convergence. BGD provides stable weight updates by processing the entire dataset at each iteration.

5.2 Logistic Regression with Stochastic Gradient Descent (SGD)

- **Parameters**: `loss='log_loss', max_iter=1000, tol=1e-3, class_weight='balanced', random_state=42`.

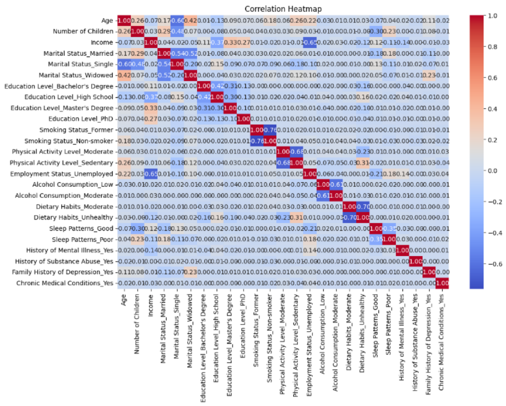


Figure 4: Positive correlations (in red) indicate that as one factor increases, so does the other, such as income and employment or physical activity and good sleep. Negative correlations (in blue) show an inverse relationship, where an increase in one factor leads to a decrease in another, such as unemployment with income or poor sleep patterns with physical activity and healthy habits.

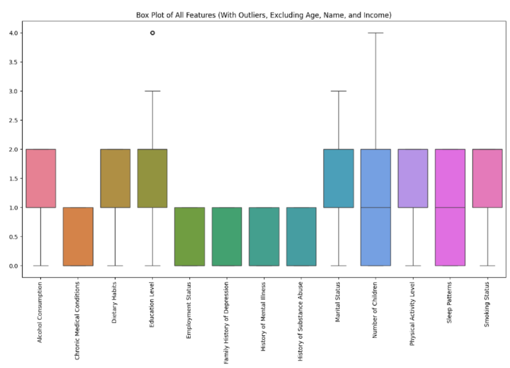


Figure 5: This boxplot detects no outliers in the data

- Faster but noisier compared to BGD, with updates on individual samples. Used to explore faster training possibilities on scaled data.

5.3 Ridge Logistic Regression with Recursive Feature Elimination (RFE)

- **Parameters**: `penalty='l2'`, `n_features_to_select=5`, `max_iter=1000`, `class_weight='balanced'`, `random_state=42`.
- RFE selects top features, and L2 regularization penalizes large coefficients to prevent overfitting.

5.4 Principal Component Analysis (PCA)

- **Parameters**: `n_components=10`.
- Reduced dimensionality while retaining most variance (explained variance ratio). Applied to simplify models and improve performance.

5.5 Optimal Threshold Tuning

- Calculated the optimal threshold to maximize the difference between TPR and FPR, improving classification for imbalanced data.

5.6 Random Forest (Tuned)

- **Parameters**: `n_estimators=200`, `max_depth=20`, `min_samples_split=5`, `min_samples_leaf=2`, `class_weight='balanced'`.
- Tuned to optimize performance with hyperparameter adjustments for better generalization.

5.7 Decision Tree Classifier

- **Parameters**: `n_estimators=200`, `max_depth=20`, `min_samples_split=5`, `min_samples_leaf=2`, `class_weight='balanced'`.
- Trained using GridSearchCV on SMOTE-balanced data. Effective at capturing non-linear patterns with controlled overfitting.

5.8 Naive Bayes (Gaussian and Bernoulli)

- **Gaussian Naive Bayes**: Assumes features follow a normal distribution, suitable for continuous data.
- **Bernoulli Naive Bayes**: Trained on binary-transformed features (values 0).
- Both models took advantage of feature independence.

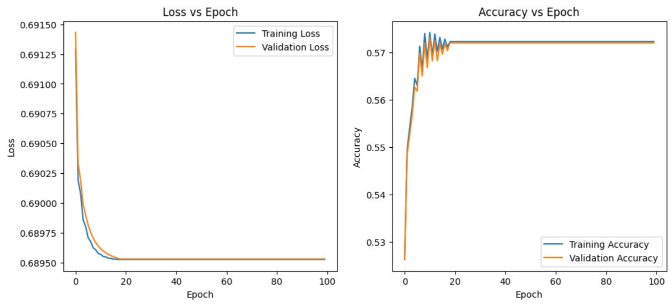
5.9 MLP Classifier (Neural Network) with Multiple Activation Functions

A classifier (MLP) was applied to assess the impact of various activation functions including **'identity', 'logistic', 'tanh', and 'relu'** on model performance

- **Parameters**: Two hidden layers with 100 and 50 neurons, 'adam' solver, a maximum of 300 iterations, early stopping enabled with no improvement stopping after 10 iterations, ensuring the model avoids overfitting.

6 Results

Plot Loss vs epoch ad accuracy vs epoch on logistics regression



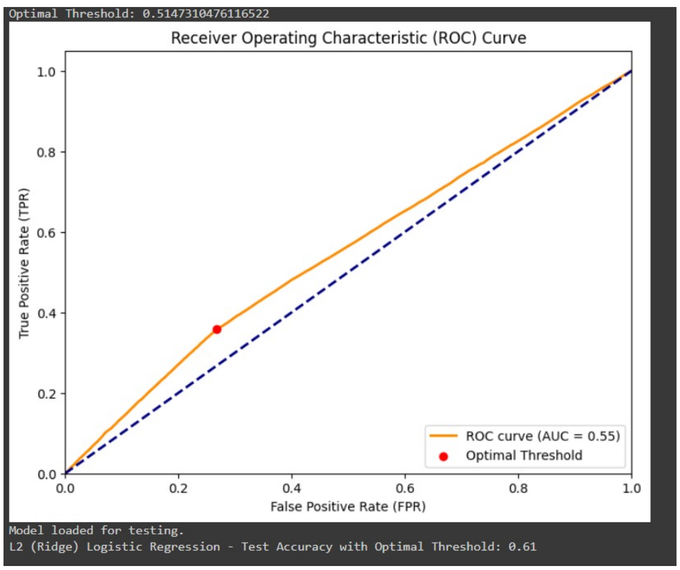


Figure 6: Logistic Regression with regulation with l2 (Ridge) by finding Optimal Threshold value 0.512 and finding the ROC curve

7 Performance Insights & Comparison Analysis

7.1 Logistic Regression with SMOTE and Optimal Threshold vs. Basic Logistic Regression

- Logistic Regression with SMOTE and threshold tuning (60%) outperforms the basic version (57%) by addressing class imbalance and refining decision boundaries, improving the separation between classes.

Machine Learning Technique	Accuracy(%)
Logistic Regression with Batch Gradient Descent (BGD)	57
Logistic Regression with Stochastic Gradient Descent (SGD)	52
L2 (Ridge) Logistic Regression with Recursive Feature Elimination (RFE)	57
Logistic Regression with PCA	54
Logistic Regression with SMOTE and OPTIMAL THRESHOLD	60
Random Forest (Default parameters)	59
Random Forest (Tuned)	61
Decision Tree Classifier [CRITERIAN='GINI']	57
Decision Tree Classifier [CRITERIAN='ENTROPY']	63
Naive Bayes (Gaussian)	67
Naive Bayes (Bernoulli)	67
MLP Classifier('IDENTITY')	58
MLP Classifier('LOGISTIC','TANH','RELU')	66

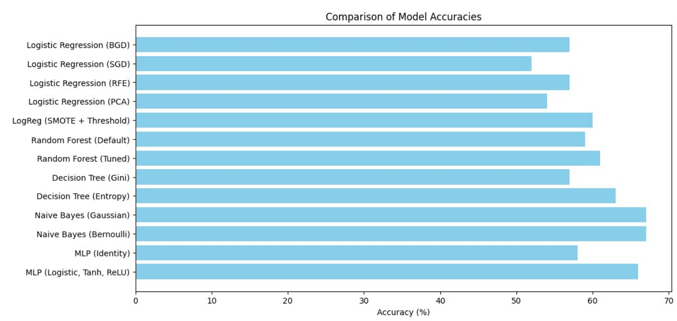


Figure 7: Plot the model which is shown above in the table

7.2 Tuned Random Forest vs. Default Random Fores

- The tuned **Random Forest (61%)** performs better than the default version (59%) due to optimized hyperparameters (like max_depth and n_estimators), reducing overfitting and improving generalization.

7.3 Decision Tree (Entropy) vs. Decision Tree (Gini)

- The entropy-based Decision Tree (63%) achieves higher accuracy compared to Gini (57%) because it captures finer information gains during splits, enhancing class separation.

7.4 Naive Bayes vs. Logistic Regression

- Naive Bayes (67%) performs better than Logistic Regression (57%) by efficiently handling the dataset's feature independence and class imbalance, making it well-suited for binary classification tasks.

7.5 MLP with Non-linear Activation vs. Identity Activation

- MLP with non-linear activations (66-67%) outperforms identity activation (58%) as nonlinear functions like ReLU, Tanh, and Logistic allow the network to learn complex patterns, improving classification performance.

8 Conclusion

Naive Bayes (67%) performed best due to the dataset's alignment with the independence assumption and SMOTE balancing. **Decision Tree (63%)** captured non-linear patterns with well-tuned hyperparameters, while **Random Forest (61%)** handled feature interactions but struggled with variance. **MLP (66-67%)** utilized non-linear activations to learn complex patterns effectively. **Logistic Regression with SMOTE and threshold tuning (60%)** improved classification but fell short compared to non-linear models. Overall, tuning, regularization, and class balancing were key to optimizing performance.

9 Individual Contribution

- **Satyam**: Decision Tree, MLP
- **Aditya**: Logistic Regression ,Decision Tree
- **Suyash**: Random Forest, Model Evaluation
- **Vickey**: Naive Bayes, Data Preprocessing
- **Satyam Pandey**: Documentation and EDA

10 Future Work

In future, we will apply SVM for better boundaries and ensemble methods like AdaBoost, XGBoost, and Voting Classifier to enhance accuracy and reduce bias. These techniques will help optimize predictions and handle complex patterns effectively.

References

[1] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. A comparative study of different classifiers for detecting depression from spontaneous speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8022–8026, 2013. 1

[2] Nikhil Chauhan and Divya Rani Swati. Depression detection using ml. 1

[3] Sergi G Costafreda, Carlton Chu, John Ashburner, and Cynthia HY Fu. Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PloS one*, 4(7):e6353, 2009. 1